

Indirect Bibliometrics by Complex Network Analysis

J. Raimbault^{1,2}

¹Géographie-cités (UMR 8504 CNRS)

²LVMT (UMR-T 9403 IFSTTAR)

Cybergeo : 20 ans déjà !

Jeudi 26 mai 2016

Context

"You are what you cite" : **Which disciplines populate the scientific neighborhood of cybergegeo ? Are they different from the ones obtained through article content (POC) and declared contents (HC) analysis ?**

→ Important for editorial policy : interdisciplinarity and Open Science

→ Semi-qualitative approach, against purely quantitative bibliometrics harmful to humanities

Objective

Research question : *How does the combination of a citation network approach with a semantic analysis unveil disciplinary context of the journal ?*

→ Hypernetwork methodology : superposition of a citation network with a semantic network, in the spirit of a transversal approach

→ Data difficult to access : database to construct

Data collection

Cybergeo data : journal production base

→ Structuration and Consolidation

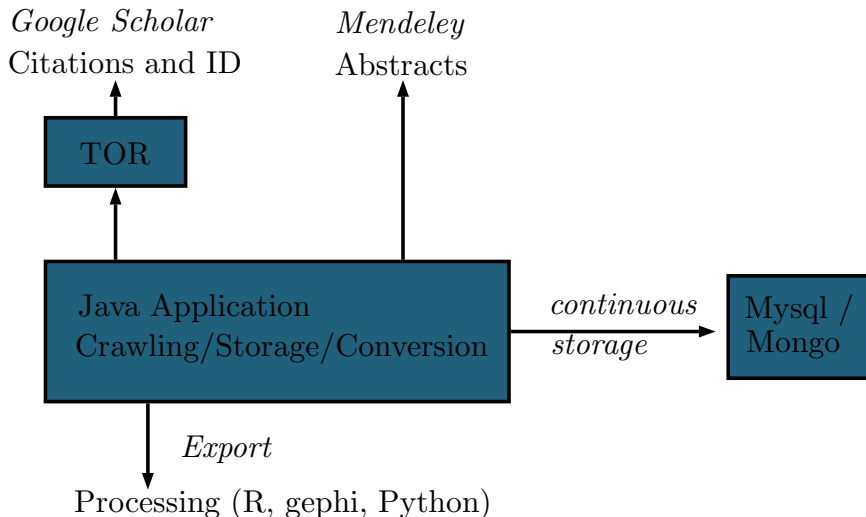
Citation data : cybergeo not indexed by “classical” bases (such as Web of Science[©], which are furthermore not open)

→ Google Scholar[©] crawling, using “*cited by*” option [Noruzi, 2005] to reconstruct citation network

Text data : need abstracts for all linked articles

→ use of Mendeley API [Mendeley, 2015] (free but not open)

Data Collection Architecture

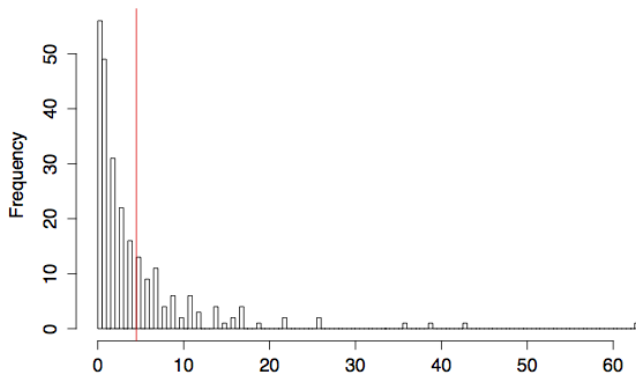


Network Properties

→ $\simeq 947$ cybergegeo articles can be studied, among $\simeq 1200$

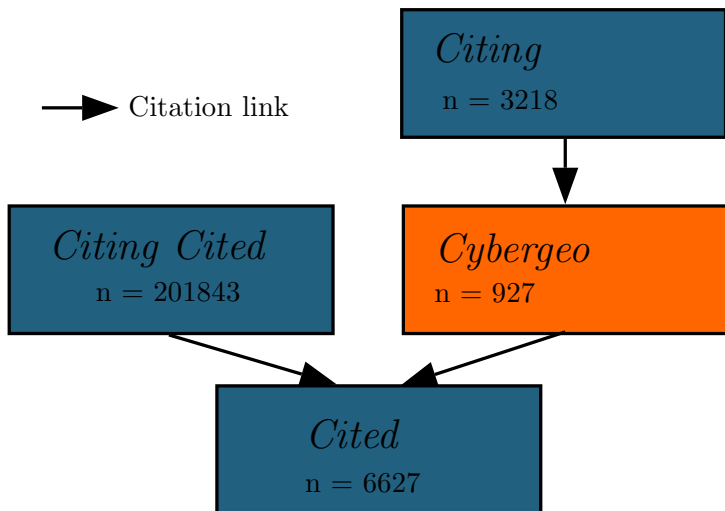
→ 418670 Nodes et 570352 Links ; Diameter : 9 ; Density : $3.25E-6$; average degree : 2.724284

Degree distribution, mean (impact factor) = 3.18

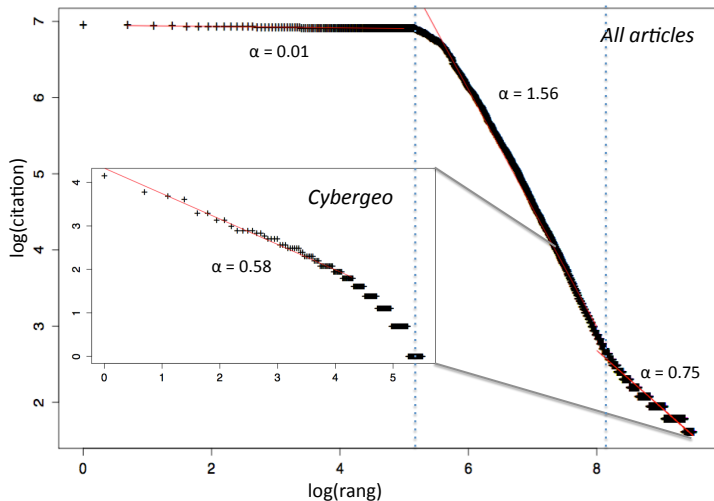


The stationary integrated impact factor, estimated as average citation count, means that a cybergegeo paper gets at least 3 citations in its lifetime

Citation Network Structure



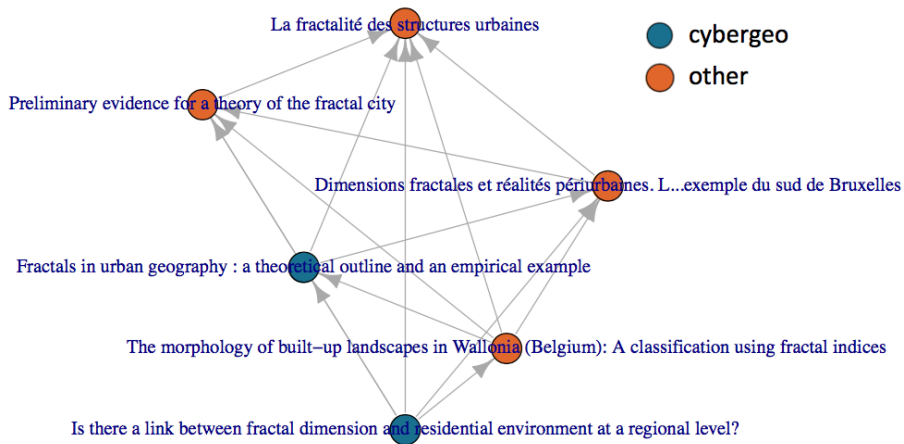
Hierarchy in citations



Superposition of different hierarchical citation regimes

Cliques

Complete subgraphs reveal strong affiliation patterns



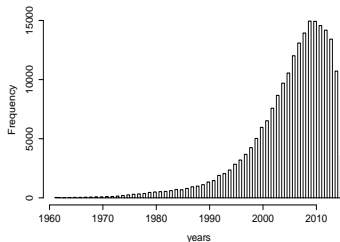
Semantic Network

Semantic Data : Collection of abstract/date/authors/keywords for the 400000 references via Mendeley API → ~ 215000 references with full data.

Summary Statistics

Language : English 206607, French 4109, Spanish 2029, German 892, Portuguese 891, Dutch 124, others 182

Yearly count



Keywords Extraction

Text-mining in python with nltk [Bird, 2006], method adapted from [Chavalarias and

- Language detection using *stop-words*
- Parsing and tokenizing / pos-tagging (word functions) / stemming done differently depending on language :
 - ▶ English : nltk built-in pos-tagger, combined to a PorterStemmer
 - ▶ French or other : use of TreeTagger [Schmid, 1994]
- Selection of potential *n-grams* (with $1 \leq n \leq 4$) : English $\cap\{NN \cup VBG \cup JJ\}$; French $\cap\{NOM \cup ADJ\}$
- Database insertion for instantaneous utilisation (10j \rightarrow 2min)
- Estimation of *n-grams* relevance, following co-occurrences statistical distribution

Construction of Semantic Network

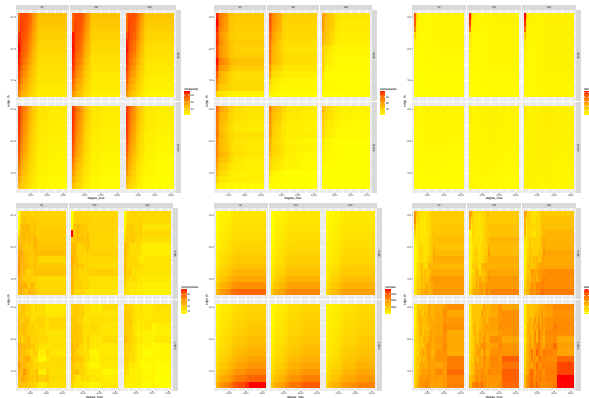
- **Nodes** : Keywords with largest relevance
- **Links** : Weighted co-occurrences
- Manual suppression of parasite words (e.g. : copyright statements !)
- Low weight link filtering
- Suppression of *hubs* (ex. model, space, structure, process) that suppress community structure
- Community detection by greedy modularity maximization (Louvain method [Blondel et al., 2008])

Parameters influence

Importance of fine tuning :

→ Sensitivity of models **and** data analysis to parameters. Systematic exploration mandatory, via OpenMole for example.

→ Place of expert decision-making : no qualitative-quantitative dichotomy



Multi-criteria optimization (modularity, size, balance) on network construction parameters

Obtained disciplines

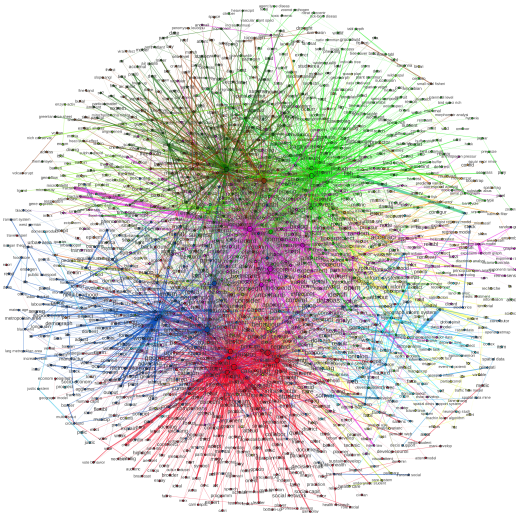
Communities obtained with $\theta_V = 1200, \theta_E = 50$

- Political sciences/critical geography (535) : decision-mak, polit ideolog, democraci, stakehold, neoliber
- Biogeography (394) : plant densiti, wood, wetland, riparian veget
- Economic geography (343) : popul growth, transact cost, socio-econom, household incom
- Environment/climate (309) : ice sheet, stratospher, air pollut, climat model
- Complex systems (283) : scale-fre, multifract, agent-bas model, self-organ
- Physical geography (203) : sedimentari, digit elev model, geolog, river delta
- Spatial analysis (175) : spatial analysi, princip compon analysi, heteroscedast, factor analysi

Obtained disciplines (continued)

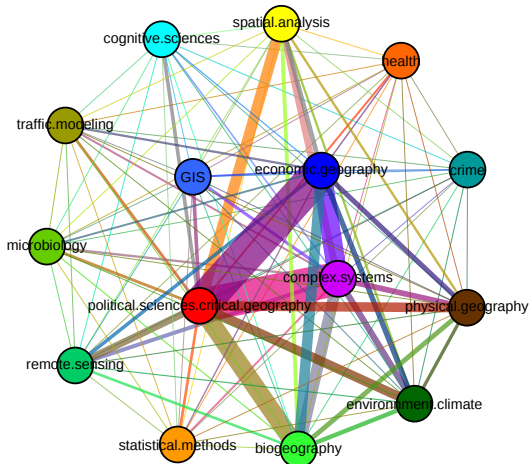
- Microbiology (118) : chromosom, phylogenet, borrelia
- Statistical methods (88) : logist regress, classifi, kalman filter, sampl size
- Cognitive sciences (81) : semant memori, retrospect, neuroimag
- GIS (75) : geograph inform scienc, softwar design, volunt geograph inform, spatial decis support
- Traffic modeling (63) : simul model, lane chang, traffic flow, crowd behavior
- Health (52) : epidem, vaccin strategi, acut respiratori syndrom, hospit
- Remote sensing (48) : land-cov, landsat imag, lulc
- Crime (17) : crimin justic system, social disorgan, crime

Network



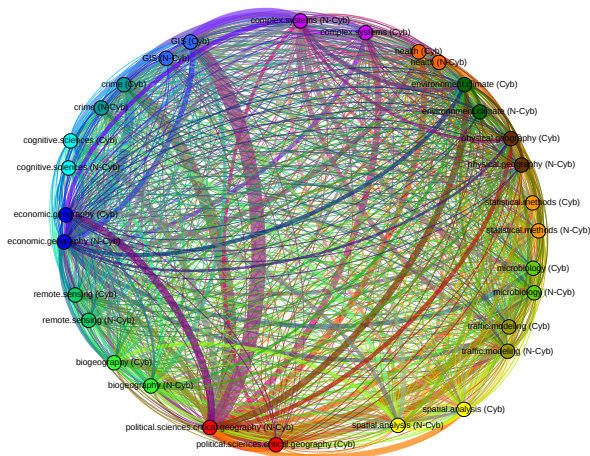
political sciences/critical geography	(19,92 %)
biogeography	(16,24 %)
economic geography	(13,31 %)
complex systems	(11,35 %)
environment/climate	(8,57 %)
physical geography	(7,82 %)
spatial analysis	(5,94 %)
microbiology	(3,83 %)
cognitive sciences	(2,93 %)
statistical methods	(2,86 %)
GIS	(2,48 %)
traffic modeling	(1,8 %)
remote sensing	(1,28 %)
health	(1,13 %)
crime	(0,53 %)

Interdisciscplinary



Synthetic representation of disciplines. Link strength gives the probability for two disciplines to jointly appear in a paper

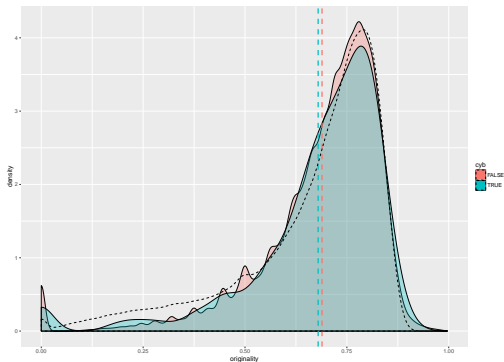
Citation interdisciplinarity



Citation flows between disciplines (directed links to be read in anti-trigonometric sense) reveal citation level interdisciplinarity

Article-level interdisciplinarity

An article has a proportion of keywords in each discipline, which can be understood as probabilities (p_i).
Interdisciplinarity index defined as $i = 1 - \sum p_i^2$.

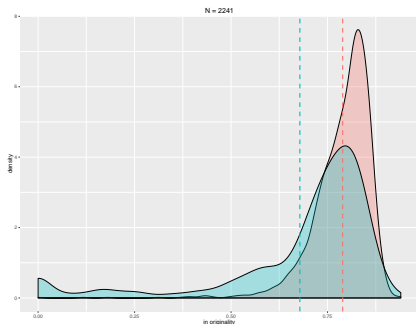


certain
 zone
 modern
 centr
 diffus
 hierarchi
 declin
 modern
 coastal
 presenc
 busi
 diffus
 compani
 diffus
 hierarch
 diffus

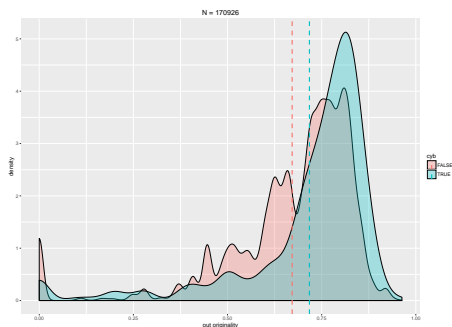
Distribution of article interdisciplinarity (null model in dotted line).

Citation interdisciplinarity

Citation interdisciplinarity defined the same way, based on probabilities to cite or be cited by a discipline.

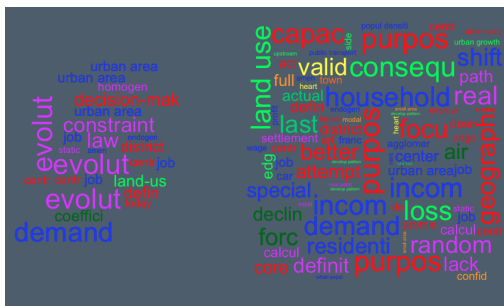


Citing article interdisciplinarity distribution. Cybergeo papers are cited by less different disciplines, what can be explained by their young age.



Cited article interdisciplinarity distribution. Distribution for all articles is directly shaped by citation network structure.

On CybergeonNetworks : Article level citation and semantic exploration ; semantic network exploration



Conclusion

- A very rich scientific environment and a confirmed interdisciplinarity
- Approach to be combined with other classifications (thematic (POC), keywords (HC), geographical (CC)) to unveil patterns in geographical practices around the journal
- Generic method that can be applied to any network whose nodes have a textual description



1996-2016 : 20 ans de cybergegeo

1996-2016 : 20 years of cybergegeo

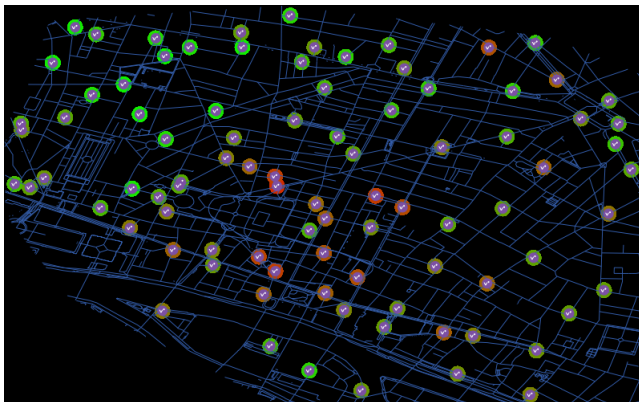
Reserve Slides

Reserve Slides

Data Collection

Crawling of semi-open data : examples in geography

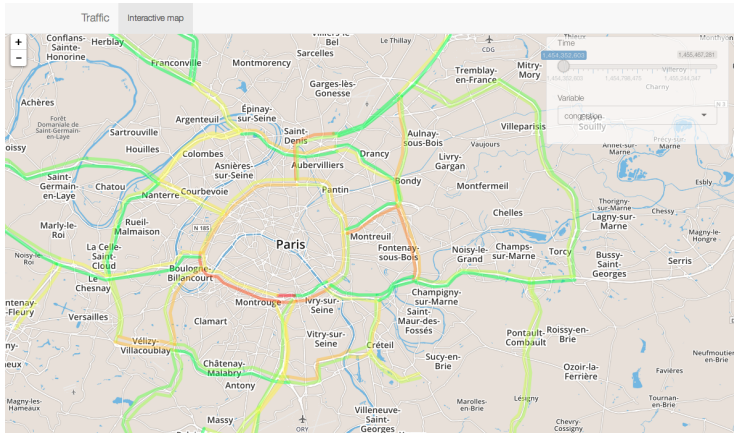
Mobility data : bike-sharing docking stations status (API)
[Raimbault, 2015]



Data Collection

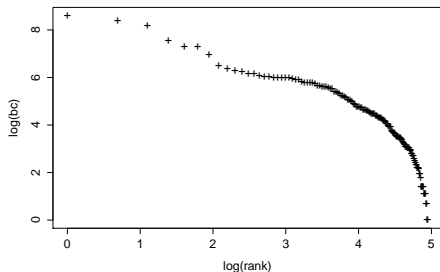
Exemples in geography (continued)

Road traffic : collect of *syta* data (no API : *scrapping* is necessary)

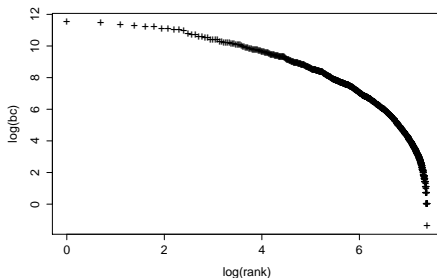


Centrality (citation)

rank-betweenness-centrality (cybergeo)



rank-betweenness-centrality

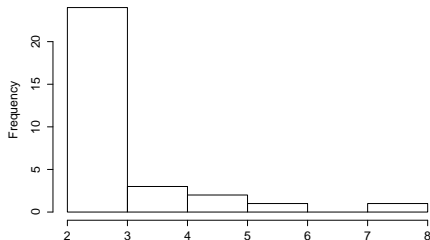


Weak centralities (rq : impossibility of having strong clusters because of temporal causality). Left : Cybergeo ; Right : Whole Network

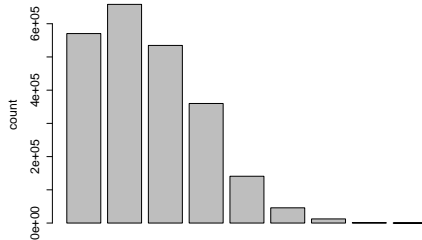
Clustering (citation)

Giant component : more than 99% of nodes.

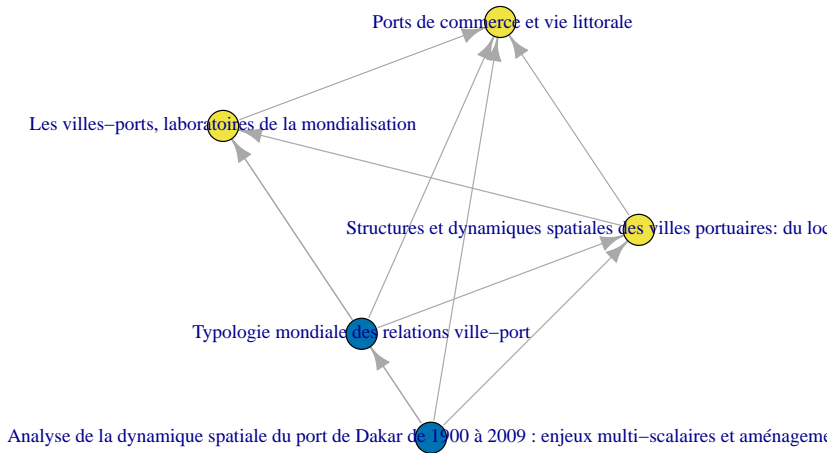
Weak clusters size without giant component



path length distribution



Cliques(citation)



29 / 33

Relevance estimation

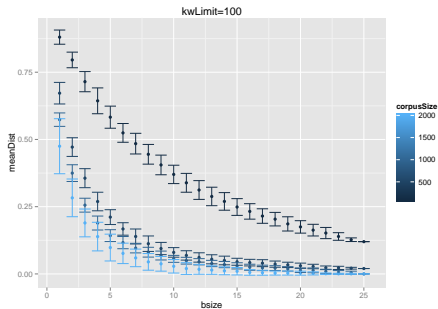
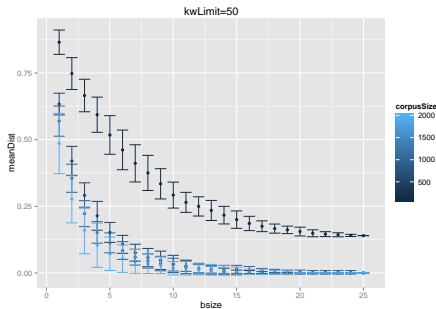
Relevance estimation via statistical distribution of co-occurrences (χ^2 score) : *termhood* defined, with M_{ij} number of articles where i et j appear simultaneously :

$$t_i = \sum_{j \neq i} \frac{(M_{ij} - \sum_k M_{ik} \sum_k M_{jk})^2}{\sum_k M_{ik} \sum_k M_{jk}}$$

in $\Theta(\sum_i N_i^2)$ (N_i abstract sizes) : difficult on a corpus where $\sum_i N_i^2 \simeq N < N_i >^2 \simeq 8 \cdot 10^7$

Relevance estimation

Bootstrap estimation tests (performance)



References I



Bird, S. (2006).

Nltk: the natural language toolkit.

In Proceedings of the COLING/ACL on Interactive presentation sessions, pages 69–72. Association for Computational Linguistics.



Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008).

Fast unfolding of communities in large networks.

Journal of statistical mechanics: theory and experiment, 2008(10):P10008.



Chavalarias, D. and Cointet, J.-P. (2013).

Phylomemetic patterns in science evolution—the rise and fall of scientific fields.

Plos One, 8(2):e54847.



Mendeley (2015).

Mendeley reference manager.

<http://www.mendeley.com/>.

References II



Noruzi, A. (2005).

Google scholar: The new generation of citation indexes.

Libri, 55(4):170–180.



Raimbault, J. (2015).

User-based solutions for increasing level of service in bike-sharing transportation systems.

In Complex Systems Design & Management, pages 31–44. Springer.



Schmid, H. (1994).

Probabilistic part-of-speech tagging using decision trees.

In Proceedings of the international conference on new methods in language processing, volume 12, pages 44–49. Citeseer.